HORIZON 2020

Digital Security: Cybersecurity, Privacy and Trust
H2020-DS-2015-1

DS-04-2015 Information driven Cyber Security Management
Grant n° 700176

Secure Information Sharing Sensor Delivery event Network[†]

Deliverable D4.1 version 1.0:

Data exchange interfaces specification

Abstract: This document specifies the public interfaces provided by the SISSDEN platform. These interfaces are devised for different types of Actors external to the SISSDEN project, including the general public and individuals acting on behalf of their organisations (research, industry, CERTs, LEAs, etc). This deliverable presents the part of the specification efforts carried out in the task T4.4 dealing with the public interfaces. The private SISSDEN interfaces and APIs are also specified in this task but are presented in the private internal SISSDEN architecture specification document (initial version delivered as D3.3). *(cont.)*

Contractual Date of Delivery	January 31st, 2017
Actual Date of Delivery	February 28th, 2017
Deliverable Security Class	Public
Editor	Montimage (MI)
Contributors	Involved partners: MI, NASK, CYBE, EXYS, DTAG, SHAD
Internal Reviewers	Angelo Consoli (EXYS), Krzysztof Lasota (NASK)
Quality Assurance	Adam Kozakiewicz (NASK), Paweł Pawliński (NASK)

[†] The research leading to these results has received funding from the European Union Horizon 2020 Programme (H2020-DS-2015-1) under grant agreement n° 700176.

Abstract (cont.):The interfaces are derived from the SISSDEN requirements and use cases but will receive periodic updates throughout the lifetime of the project, to allow subsequent changes, improvements and optimizations to be included.

The interfaces described here will allow users to access publicly available data (e.g., website, news) and enable access to other types of data requiring subscription and validation by SISSDEN to external parties via the SISSDEN data sharing platform. The data includes remediation reports, metrics, aggregated statistics, and curated reference datasets, but also allows users to provide feedback and control their privacy.

The *SISSDEN* consortium consists of:

Naukowa i Akademicka Sieć Komputerowa	Coordinator	Poland
Montimage EURL	Principal Contractor	France
CyberDefcon Limited	Principal Contractor	United Kingdom
Universitaet des Saarlandes	Principal Contractor	Germany
Deutsche Telekom AG	Principal Contractor	Germany
Eclexys SAGL	Principal Contractor	Switzerland
Poste Italiane – Sozieta per Azioni	Principal Contractor	Italy
Stichting the Shadowserver Foundation Europe	Principal Contractor	Netherlands

Table of Contents

1 Introduction

1.1 Aim of the document

The SISSDEN platform will provide a number of interfaces and APIs that will be made available to the public and/or external partners so that they can access and contribute data to the SISSDEN project. Interfaces will be implemented by web sites, an email gateway, Customer Portal, metric dashboard, etc. These public facing systems will include mechanisms to communicate with the consortium, sign up to request free of charge reports, gain access to the curated reference data set, provide customer feedback, and manage opt in/out and data privacy issues.

The scope of this document is limited to the interfaces and APIs that will be made public and concern mainly human-machine interactions. Specification of private internal interfaces and APIs will be included in the private internal architecture document (initial version was delivered as D3.3) and will concern mainly machine-to-machine interactions.

The first release of this document (in M9+) is mainly a high level specification of these interfaces and a justification of the choices made. It first identifies the interfaces required by the different actors involved and supported data processes. It then provides a first specification of public data sharing interfaces, APIs and formats.

This document is a living document, and will be updated and maintained throughout the lifetime of the project. The interfaces and APIs may well change and evolve over time, as the SISSDEN Partners learn more during the development and operation of the platform. It will follow the evolution of the private internal SISSDEN technical architecture, specified initially in D3.3 and included in future deliverables D3.4 and D6.7, and keep a coherent description of the SISSDEN public data exchange interfaces at all times. The final version of this document will represent the interfaces and APIs that will be made publicly available by the end of the project.

1.2 Work done so far

Inputs for the definition of the public interfaces and APIs is based on the work performed by the SISSDEN consortium members with respect to the specification of the initial technical architecture (D3.3), of which the initial draft was already submitted to the EU(M9). Note that D3.3, like this deliverable, is also a live document subject to change.

D4.1 also includes work performed by the SISSDEN consortium members so far, including:

1. The initial SISSDEN proposal (M0),
2. The existing works specified in the consortium members' backgrounds (as defined in the signed Consortium Agreement (M0)),
3. The Use Cases and Requirements D3.1 (M6),
4. The survey of potential data sources D3.2 (M6),
5. The guidelines for data handling D1.4 (M6),
6. The preliminary legal requirements D2.2 (M9),
7. The face to face consortium meetings in Warsaw (M1) and Rome (M9),
8. The conclusions of other discussions, e.g., during the teleconferences and the technical meeting in Bratislava (M6).

1.3 Structure of the document

The document starts with Section 2, which presents a résumé of the project's concepts and of the benefits it will bring to the public.

Section 3 introduces SISSDEN related stakeholders and actors, and their different roles and requirements. It includes a simplified overview of the initial SISSDEN architecture and basic concepts, highlighting public interfaces.

Section 4 identifies and describes the different types of actor and data accesses that will be implemented. It identifies the data formats, protocols used and the interactions involved, mainly focusing on human-machine interactions.

Section 5 describes the data feeds that will be provided to the stakeholders, including the free of charge victim remediation reports.

Section 6 describes the security access controls that will be put in place and discusses some of the privacy and anonymization issues which are being addressed in more detail in deliverables D1.4 (internal) and D2.2.

Finally, Appendices describe one of the data formats in greater detail and provide scaled-up, more readable versions of key diagrams.

2 Introduction of the concepts of the SISSDEN project and its benefits to the public

D4.1 is the first technical deliverable made public by the SISSDEN consortium. This section is intended to provide to new readers that are unfamiliar with the project concepts and goals an introduction and general overview of the developed platform.

2.1 Project overview

SISSDEN is a Horizon 2020 project aimed at improving the cybersecurity posture of EU entities and end users through the development of situational awareness and sharing of actionable information. It builds on the experience of The Shadowserver Foundation, a non-profit organization well known in the security community for its efforts in mitigation of botnet and malware propagation, free of charge victim notification services, and close collaboration with Law Enforcement Agencies, national CERTs, and network providers.

The core of SISSDEN is a new worldwide sensor network, which will be deployed and operated by the project consortium from within the EU. Passive threat data collection mechanisms will be complemented by behavioural analysis of malware and multiple external data sources. Actionable information produced by SISSDEN will be used for the purposes of no-cost victim notification and remediation via organizations such as National CERTs, ISPs, hosting providers and Law Enforcement Agencies such as Europol's European Cybercrime Center (EC3). SISSDEN will especially benefit SMEs and citizens, which do not have the capability to resist threats alone, allowing them to participate in this global effort, and profit from the improved information processing, analysis and exchange of security intelligence, to effectively prevent and counter security breaches. The main goal of the project is creation of multiple high-quality feeds of actionable security information that will be used for remediation purposes and for proactive tightening of computer defences.

This will be achieved through development and deployment of a distributed sensor network based on state-of-the art honeypot/darknet technologies and creation of a new EU based high-throughput data processing centre. SISSDEN will provide in-depth analytics on the collected data and develop metrics that will be used to establish the scale of most important security issues in the EU, and impact of the project itself. Finally, a curated reference data set will be created and published to provide a high-value resource for vetted academic and private industry security researchers.

2.2 Project objectives

The SISSDEN project has specified the following eight objectives as core goals of its operation:

1. **Create a large distributed sensor network.** Over 100 passive sensors based on current and beyond state-of-the-art honeypot and darknet technologies will be deployed in multiple organizations, including all 28 EU member states and 6 candidate countries. These sensors will be used to observe malicious activities on an unprecedented scale, without intercepting any legitimate traffic.
2. **Advancements in attack detection.** New types of honeypots, darknets and probes will be deployed to detect, analyse and alert on types of attacks not

widely detected today, such as reflective DDoS amplification or attacks against Internet of Things (IoT) devices, which are expected to increase significantly in the coming years as a range of new network-centric technologies are embraced by consumers and SMEs globally.

3. **Advancements in malware analysis and botnet tracking.** The large sensor network will be augmented by an innovative new generation of enhanced sandbox technologies designed for long running monitoring of malware specimen execution and behavioural clustering, to provide even more information on current threats.

4. **Improving the fight against botnets.** Sensor and sandbox data collected will be used for detailed studies of botnet infrastructures. Long term observation of multiple families of current botnets will support anti-botnet research and law enforcement activities. Output will closely align with existing European anti-botnet and anti-cybercrime strategies, as well as providing support to proven strong LEA partnerships, such as with Europol EC3.

5. **Collect, store, analyse and reliably process Internet scale security data sets.** The inherent challenges of building and continuously operating reliable data collection, storage, exchange, analysis and reporting systems at high volumes will be solved by multiple innovations in sensor and backend packaging, deployment, integration and data searching, based on consortium members' extensive experience with "big data" approaches, high volume transactional and non-relational data systems.

6. **Share high-quality actionable information on a large scale.** SISSDEN will produce large amounts of intelligence on current threats and all of it will be shared with stakeholders and the larger community, at no cost to them, for the purposes of remediation or for early warning. The project will distribute high-quality data feeds to the majority of the National CERTs in Europe, as well as worldwide, along with Law Enforcement Agencies, Internet providers, network owners and other vetted organisations fighting to defend their networks, SME customers, EU citizens and Internet Users against continuous attacks.

7. **Provide objective situational awareness through metrics.** The consortium will have access to huge amounts of high-quality data on cyber threats: primarily obtained by the sensor network but also contributed by the members of the consortium. This unprecedented visibility will enable metrics developed as part of SISSDEN to offer a truly objective, non-vendor biased overview of the threat landscape in the EU and individual member states.

8. **Create and publish a large scale curated reference data set.** A significant subset of the data produced by SISSDEN will be made available to vetted researchers and academia, addressing the clear and urgent need for large scale, high quality, recent security datasets in order to improve or test defensive solutions. This should become a valuable new resource for powering security research excellence in Europe.

2.3 Functionality of the project

The initial months of SISSDEN project activity (M1-M6) have resulted in a set of detailed use cases and functional requirements being developed internally by consortium partners (D3.1). Potential external data enrichment sources have also be identified (D3.2), as well as guidelines for data handling (D1.4) and preliminary legal requirements (D2.2).

These documents allow conceptual sets of required functions to first be defined at a high level in section 3, and to then be broken down into more logical detail in the sections 4 and 5. Implementation of the initial SISSDEN public data interfaces and APIs will include the following conceptual elements:

Large-scale security data collection. This is the core of the SISSDEN project activity, enabling all other conceptual elements. The task is both well-defined and complex, as apart from the inherent challenge of handling large amounts of data, the project must deal with a very wide array of data sources, both heterogeneous and geographically distributed. Inclusion of multiple types of existing honeypots is only a starting point; the project is expected to deliver new beyond state of the art honeypot technologies. The project will deploy such honeypots across the EU but also expects them to be deployed by interested third parties. Various other types of threat data feeds are also planned for integration, most of them made available by third parties – a survey of such data sources was presented in the deliverable D3.2 “Existing data sources catalogue”.

Beyond state of the art data analysis capability. The wealth of information available in the system enables advanced and innovative analysis methods to be developed and used. This includes new methods of analysing honeypot and darknet data, new methods for malware analysis, botnet and malware tracking, such as advanced sandboxing approaches including long-term execution. The various methods developed within the project will be implemented as analysis modules available from the collaborative analytical platform which will be instrumental in future research based on the project outcomes.

Extraction and delivery of actionable information. A core function of the SISSDEN system is its ability to provide such information to parties responsible for the affected networks (e.g. network providers, national CERTs, etc.) in order to enable effective defence against Internet threats. Other information sharing is also possible, e.g. providing the information on illegal activity to Law Enforcement Agencies.

Global situational awareness. The system processes significant amounts of data from sources distributed around the whole of Europe (at least). This provides a unique global view of the state of European cyber security. A challenging but important goal of the project is to develop useful and informative metrics summarizing the observed activity. Such metrics, as well as various statistics about the collected data, can be of great value to all parties involved in the protection of European networks.

Curated reference data set. The large amount of raw data captured by the project enables the creation of a unique research resource – a large curated reference dataset, to be made available to vetted security researchers. The creation and curation of such a dataset is partially a manual process, since personally identifiable information protection rules must be strictly observed.

It should be noted that most of the functionality specified above is expected to be delivered at the highest technology readiness level (TRL 9), that is, as a fully functional system,

deployed and operating as a mature product. Only the data analysis tasks, specified as new research and not always sharing a deep grounding in pre-existing solutions, are expected to be delivered as prototypes. However, even in this case, the expected readiness level is still high – it is TRL 7, requiring the complete prototype to be fully verified in an actual operational environment. This means that the SISSDEN platform will deliver real world, tested, working data collection and data analysis solutions.

2.4 Data Sharing Requirements

The high-level description of the project makes it possible to identify the range of interfaces necessary to fulfill the project's data sharing needs. This section takes a look at the functionality specified above, identifying such high-level needs – some of them crucial, others optional.

A system of the size of the SISSDEN platform will necessarily include a large number of different interfaces between the various components, external systems and Actors. This public report focuses on public human-to-machine interfaces only. The internal or confidential interfaces are described in the architecture document – a live internal (confidential) document. The initial version of this document has already been delivered as deliverable D3.3 "Initial Technical Architecture". Updated versions of this document will form the core parts of the deliverables D3.4 "Final Technical Architecture" and D6.7 "Architecture Whitepaper".

Large-scale security data collection. This activity involves two types of interaction with third parties – the deployment of SISSDEN sensors by third parties and the inclusion of third party data feeds in SISSDEN. Neither of these activities requires a public interface. The communication between a SISSDEN sensor deployed by a third party and the system backend uses internal interfaces, just as if the sensor was deployed by SISSDEN itself. Third party data feeds provide their own interfaces. It is counterproductive to expect feed providers to implement an additional interface to be compatible with SISSDEN. Instead, SISSDEN will implement the necessary access modules. A user interface enabling third parties to declare willingness to participate in either activity in an automated way may be useful, but it is not necessary to implement the functionality.

Beyond state of the art data analysis capability. The collaborative analytical platform clearly requires a user interface for performing the available analyses. A query API enabling integration of SISSDEN analyses in more complex ad-hoc analytical scenarios is also potentially useful. However, as the analysis available in deliverable D2.2 "Preliminary Legal Requirements" shows, the raw data available in SISSDEN contain personally identifiable information (PII). Since many advanced analyses work on data potentially including such information, the results may also contain such PII. In fact, in many cases removal of PII would render the analyses to be of limited value, or even useless. Therefore, the interfaces of the collaborative analytical platform will not be public. Initially, these interfaces will only be accessible by the researchers involved in the SISSDEN project. Hence they are not covered in this public document, but will be described in the architecture document instead. Later, they may and should be made available to other, properly vetted researchers, but such access will require an appropriate legal agreement specifying the responsibility inherent in personal data processing.

Extraction and delivery of actionable information. This activity clearly includes public interfaces, as third parties must be able to register and receive the reports containing

actionable information. Other modes of data sharing possible as part of this group of functions may or may not require formal interfaces. For instance, interfacing with law enforcement regarding illegal activity is unlikely to be fully automated, as the legal acts regulating such cooperation are not unified across Europe and different national law must be followed in each case.

Global situational awareness. The metrics and statistics provided by the project must be made available through a user interface. An API for automated access is possible but not strictly required. At least for a subset of the metrics, this interface should be fully public. More detailed information, especially focusing on specific networks, etc. will need to be restricted to parties with justified interest.

Curated reference data set. The developed curated dataset must be made available to vetted researchers in some way. An interactive user interface would have limited use, as the dataset is too large to analyse manually. However, the dataset is also static, so any means of downloading it would be sufficient. Note that the dataset is not public, since it contains PII. It will be made accessible only to vetted security researchers. Interactions with the curated reference dataset are covered in the document.

Apart from the above, the amount of different services offered by SISSDEN makes it necessary to provide users with a web-based interface enabling them to, e.g., learn about SISSDEN, access the public outcomes, register for non-public outcomes, exercise their privacy rights via opt-in/opt-out mechanisms and provide feedback on the various interfaces and the provided data.

3 SISSDEN Initial Technical Architecture and Actor Roles

In section 3 we review the SISSDEN initial technical architecture (D3.3), identify the internal and external SISSDEN Actors who will utilize the system, outline the data processing needs defined in the SISSDEN Requirements and Use Cases (D3.1) and then map these to the mechanisms and formats that will be supported by the SISSDEN platform's initial technical architecture (D3.3). Human-machine interactions are primarily highlighted, with machine-machine interaction being covered in the private internal architecture document (live document, delivered as D3.3 and to be part of future deliverables).

3.1 SISSDEN Initial Technical Architecture

The figure below provides a simplified visualization of the SISSDEN initial technical architecture, described in detail in D3.3 and presented here in this public deliverable to give a new reader an understanding of how the SISSDEN platform will operate. While the public interfaces and interactions are the core of this document and described more elaborately later, we will also briefly provide an overview of the system internals, including the private interfaces.

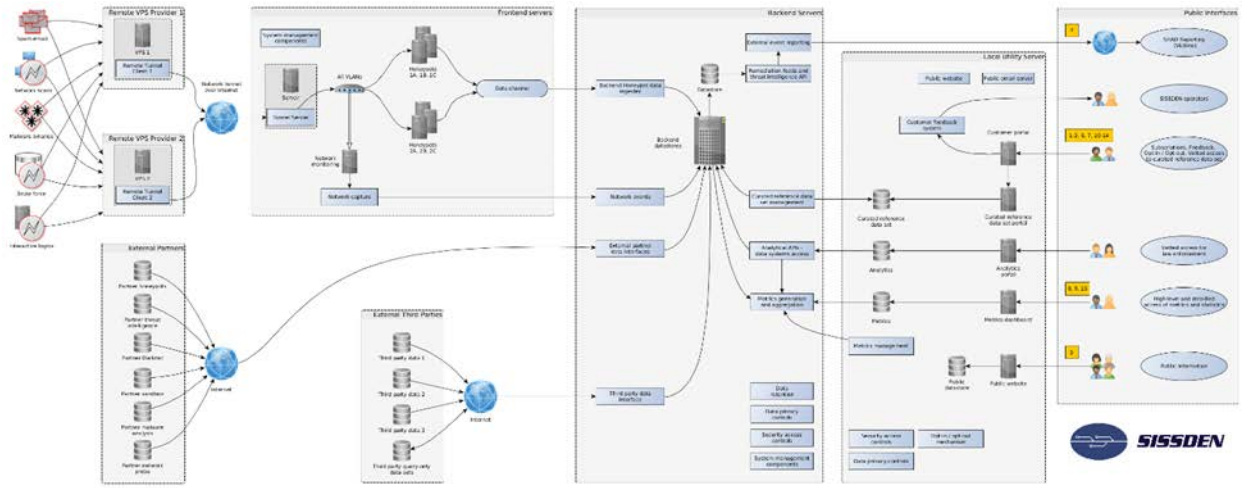


Figure 3.1: Initial SISSDEN technical architecture with highlighted public interfaces

A larger, zoomable version of this diagram is provided in the appendices at the end of this document.

As part of the project, SISSDEN is procuring a new EU datacentre and deploying a sensor network of 100+ sensor end points throughout Europe and the wider world. Components located at the EU datacentre include the Frontend Servers, Backend Servers and Utility Server pictured on the diagram. The sensor network consists of remote VPS Provider end points located at various VPS hosting providers (i.e. outside the EU datacentre), configured as transparent network tunnel endpoints forwarding traffic to the EU datacentre. SISSDEN will collect attack data such as network scans, spam email, malware binaries, brute force attacks, interactive attacker logins etc.

3.1.1 Remote endpoint sensors (VPS)

Each remote endpoint sensor will contain only the minimum amount of configuration and management capabilities required to securely participate as one end of a transparent

network tunnel. They will be configured to essentially act as a long virtual Ethernet cable between the VPS and SISSDEN's local data centre frontend. At the Frontend in the EU Datacentre, a tunnel server will terminate each transparent layer 2 Ethernet tunnel and deliver the Ethernet frames to an isolated, dedicated Virtual Local Area Network (VLAN).

3.1.2 Frontend servers

Traffic from the remote sensor endpoints will be received by multiple types of honeypot systems, implemented as VMs, running on the EU Datacentre Frontend. Each honeypot VM will emulate one or more potential vulnerabilities and collect data about attacks observed against those vulnerabilities. The honeypots will have a standard configuration and standard data collection formats enabled. Their data collection capabilities will be complemented by network packet capture components running on separate VM instances that will listen to all traffic coming to them. SISSDEN system management components will be required to centrally manage all VM configuration, orchestration and operations.

Honeypot data and data from the network capture components will be ingested into the Backend datastores located at the Backend Servers at the EU Datacentre.

3.1.3 External partner and third-party systems

The data collected by the SISSDEN sensor network will be supplemented by data from external systems operated by SISSDEN partners. These will include separate honeypot networks, darknets, sandbox and malware analysis systems, threat intelligence platforms etc. As with the sensor network, data from these systems will be ingested in various forms and stored in the Backend datastores.

Additionally, previous work in the project by partners (D3.2) has identified various candidate external third party data sources that could potentially be included as sources of enrichment for SISSDEN's own data collections.

To avoid unnecessary software development, SISSDEN will make use of and extend background partner systems, which will aggregate data from multiple sources and provide a well-defined RESTful API for accessing normalized datasets.

3.1.4 Backend Servers

Data from SISSDEN's various data collection systems will be represented in multiple formats, such as live-streamed events, log files, PCAP files, and other file format data. Most of these data types will need to be stored in their raw format in local data storage systems, at least for predetermined periods/repository size quotas, and some of the data types will also require parsing, normalization and ingesting into backend data indexes in support of free daily remediation report generation, data analytics and ad-hoc queries.

At the Backend Servers and Utility Server, SISSDEN components will be required to deploy and manage the data ingest and storage systems. This will include managing security access controls, data retention, data privacy controls etc.

3.1.5 External reporting system

One of the main purposes of the SISSDEN project is to collect Internet scale, timely security event data and make it available at no cost to vetted National CERTs, Network Owners and organizations who sign up for SISSDEN's free daily alerts.

The various sources of data collected by SISSDEN, such as honeypot and darknet data, malware analysis data, and botnet tracking information – as well as ingested external third party data sources – will be collected and stored locally in the SISSDEN backend. Each day recipients who have voluntarily signed up for free reporting will receive by email multiple reports, covering different types of potentially malicious activity detected by SISSDEN on their nominated, verified IP / ASN / CIDR addresses.

To avoid unnecessary software development, reduce costs and provide access to a large, proven, vetted reports consumer base, SISSDEN will make use of and extend Shadowserver's existing background daily reporting system. SISSDEN components will be required to process collected data and provide it to Shadowserver's external reporting system for distribution.

Details of the signup process are presented in Section 4. Initial example report formats can be found in Section 5.

3.1.6 Utility Server

Various analytics will be performed on the data collected by SISSDEN. An analytics platform will be developed, and hosted on the Utility Server. These analytics solutions will provide additional insight into threats propagating in the Internet, pooling together partner resources dedicated to the project. In addition, metrics will be applied to the collected datasets to provide improved situational awareness. They will be used as a basis on which informed decisions can be made to mitigate threats. Curated reference datasets will also be made available to vetted researchers through the Utility Server. Interactions with the above are described in more details in this document and take place through the external interfaces illustrated in the diagram (with the exception of the analytics platform, which will only be available to SISSDEN partners).

SISSDEN will present a number of systems to interact with the public and external partners. These will include a Public website (mostly containing information about the project), email communication (reports), a Customer Portal, Metrics Dashboard etc. Hosted on the Utility Server, these public facing systems will include mechanisms to communicate with the consortium, sign up to request free of charge reports, gain access to the curated reference data set, provide customer feedback, and manage opt in/out and data privacy issues. These interactions are described in more detail in the following sections of the document (the numbers on the diagram correspond to the list of interactions in Section 4). Not included is a description of access to the analytics platform, which is not publicly exposed.

3.2 SISSDEN Actors

The SISSDEN platform will have a number of internal and external Actors potentially accessing it. Each Actor role will require access to different types of data and access controls.

The table below lists the SISSDEN actors:

Role	Description and interfaces needed
System administrators	Need to administer the technical infrastructure providing the platform/infrastructure. They are not directly involved in any of the processes related to the public interfaces. They will assure that the different component systems function correctly.
Operators	Need to configure and operate the services delivered through the platform/infrastructure. They are indirectly involved in all the processes related to the public interfaces and administer the website and portals, the User profiles and subscriptions, the news, report and feedback systems.
Data providers	Data feed providers of additional data used to enrich the SISSDEN collected data. The interfaces used by them will be private and are described in D3.3.
Data consumers	Consumer of SISSDEN data. The following roles identify and describe different types of data consumers: Anonymous Users, Users, Vetted Users, Report Recipients, Approved Researchers, LEA Users, Metrics Recipients and SISSDEN Partners.
Anonymous Users	Anonymous Users are any users that are not registered for a User account on the SISSDEN website. They will have access to publicly available data only, that includes data published on the SISSDEN website, some news reports, high-level aggregated metrics, and Twitter messages. Typically these Anonymous Users are EU or non-EU citizens.
Users	Users with a valid SISSDEN account. These Users have created an account and validated it using an emailed link. Users can request access to remediation feeds, curated data, metrics or other services. They can also subscribe to general news reports concerning the project.

Role	Description and interfaces needed
Vetted Users	Users that have undergone a process of validation to obtain the authorisation to access certain types of information. Typically these Users are individuals working for or representing some entity (CERTs, LEAs, research entities, industry partners, etc.) but could also be interested EU or non-EU citizens. Vetted Users have access to customer feedback services, subscription services to request access to other data, account management services, privacy settings, and opt-in/opt-out settings. Vetted Users is a generic category that is split into more specific Vetted User roles, related to data types such as free of charge remediation feeds, metrics, reference datasets or access to law enforcement services. Additional Vetted User roles may be identified during the duration of the project and will be expanded here.
Report Recipients	Vetted Users that can receive free daily remediation reports/feeds related to their networks or constituency. This is a user identified as owning a network (e.g., based on WHOIS records), phone and other records that need to be provided to prove ownership (e.g., https://www.shadowserver.org/wiki/pmwiki.php/Involve/GetReportsOnYourNetwork). It could also include an official National CERT user, in which case an entire country is recognized as the CERT's constituency.
Approved Researchers	Vetted Users that have been successfully vetted for access to a curated reference dataset. Vetting involves a process which requires signing a terms of use or NDA (Non-Disclosure Agreement). Rules will be developed to determine the eligibility criteria, e.g., who should be a recognized researcher, or a student who has a recommendation from a supervisor.
LEA Users	Vetted Users identified as a member of a Law Enforcement Agency that in special cases requires access to the Analytics platform. Validation is performed here on a case-by-case basis. Note that the SISSDEN consortium does not view this case as public interface access.
Metrics Recipients	Vetted Users that have been successfully vetted for receiving access to detailed metrics and statistics. Typically these Users are network operators interested in the security of their network, and will be eligible to view detailed metrics about only their own network(s). The possibility of making certain advanced metrics available to non-network owners will also be considered.
SISSDEN Partners	SISSDEN Partners will need to maintain and control the platform, i.e., act as operators and administrators of the platform. For this they will have access to the services and data they will be responsible for.

Role	Description and interfaces needed
National CERTs	Primary consumers of country-level SISSDEN threat intelligence. They will typically have access to the remediation feeds at a country level (see Report Recipients), after they undergo a vetting process. Can also act as threat intelligence providers but these interfaces are private and described in the architecture document.
Internet Service Providers (ISPs), Industry partners, Telecom operators	Primary consumers of SISSDEN threat intelligence related to their internal and customer systems. They will typically have access to the remediation feeds at their own network or customer system level (see Report Recipients), after they undergo a vetting process. Can also act as threat intelligence providers but these interfaces are private and described in the architecture document.
Law Enforcement Agencies (LEAs)	Primary consumers of SISSDEN threat intelligence related to criminal investigations. Interactions with LEA is on a case by case basis. In particular, they may receive access to the Analytical platform if they need it to support their investigation (see LEA User). Can also act as threat intelligence providers but these interfaces are private and described in the architecture document.
Academic and Private Industry Researchers	Members of the research community. They will typically be granted access to the curated reference datasets if they successfully undergo the researcher vetting process (see Approved Researchers). Can also act as threat intelligence providers but these interfaces are private and described in the architecture document.
EU Citizens, Non-EU Citizens	Users that do not represent any of the other organisations described. They can access information made available through the Anonymous User role, or can subscribe to SISSDEN (create a User account) to be able to access User interfaces. They will not have access to data provided through the interfaces requiring authorization, unless they successfully undergo a vetting process and become a Vetted User.

3.3 Analysis systems

One of the major goals of the SISSDEN project is to develop a collaborative analytics portal, giving its users access to novel analysis techniques and the wealth of data collected by the system. The portal will be a valuable modern web-based tool for exploration and analysis (statistics, comparative analysis, etc.) of the large-scale data collected by the different SISSDEN system components. The system will enable collaborative analysis sessions, allowing researchers to work together and compare their results. The large-scale data analysis workflow will allow querying of heterogeneous data storages where the data collected from the SISSDEN components will be stored (log files, binary data-sets, as well as data acquired and processed from malware analysis).

The analysis system will be based on widespread open source general-purpose data processing systems as well as modern user-friendly tools, combining the strengths of interpreted languages for data analysis with new in-browser visualization libraries and collaborative capabilities. The platform brings together multiple analysis modules and offers

a common ergonomic interface, making it possible for SISSDEN Partner researchers to perform complex analytical tasks.

As the description above implies, the analytical platform will offer an advanced graphical user interface, expected to feature innovative visualization techniques supporting exploratory data analysis. However, since the analyses will work with raw data available in the system, using such data likely entails processing personal data. Automatically limiting the scope of accessible data depending on the user would not only reduce the quality of results, but also increase significantly the complexity of the platform and lower its efficiency. Therefore the interface will not be accessible to most users of the platform. Initially, the only actors with access to the analysis portal will be SISSDEN Partners. They will be able to use the interface to perform arbitrary analyses on all available data. Access to the platform could be granted to properly vetted LEA Users, but only on a case-by-case basis. Inclusion of other types of Vetted Users (using a very strict form of the vetting process) may be considered in the future, but is not planned at this stage.

In conclusion, the user interface of the collaborative analysis platform is not public and its details will be elaborated in the (confidential) architecture document.

4 Public Interface Details

Having described the high level SISSDEN initial technical architecture and associated Actors, we now describe each of the SISSDEN public interfaces in more detail.

4.1 Customer Portal Account Sign Up

Actor(s): Anonymous User.

Interface type(s): Web interface.

Interface location: Customer Portal.

Format(s): Web form, Verification email.

Interaction: An Anonymous User can visit the public SISSDEN Customer Portal component (from link in Public Website) and sign up for a free SISSDEN account. Submitted User account information will be automatically validated by the public website component.

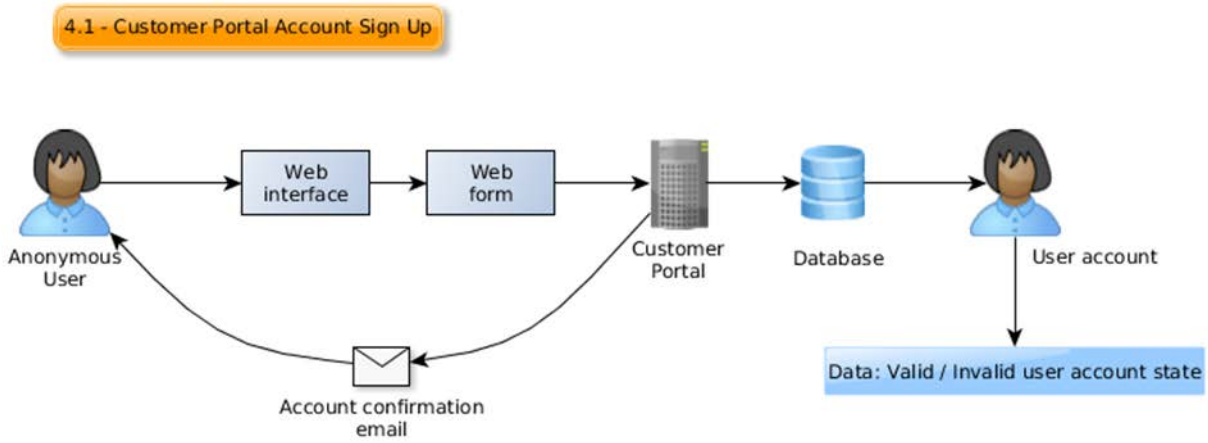
Data type(s): Creation of User account (account name, password, email address), validation of User email address (email with confirmation link), valid/invalid User account status.

Actor state change: Anonymous User -> User.

Security controls: HTTPS/SSL website, Username/password, TLS email (or other).

Data privacy: User data stored only on SISSDEN EU data centre web server, with privacy controls and opt-in/opt-out controls available.

Diagram:



4.2 Free Remediation Report Sign Up

Actor(s): User.

Interface type(s): Web interface.

Interface location: Customer Portal.

Format(s): Web form, Verification email/ticket.

Interaction: A User with an already validated User account visits the public SISSDEN website component and signs up to request access to free of charge daily remediation reports about their networks. Details of the User's network space (ASN/IP/CIDR) will be requested. The network space information collected will be used to manually validate User's ownership of the requested network space and manually verify their eligibility to receive free daily remediation reports (done by SISSDEN Partners, i.e. SHAD). This User becomes a Vetted User for receiving free remediation reports, i.e. a Report Recipient.

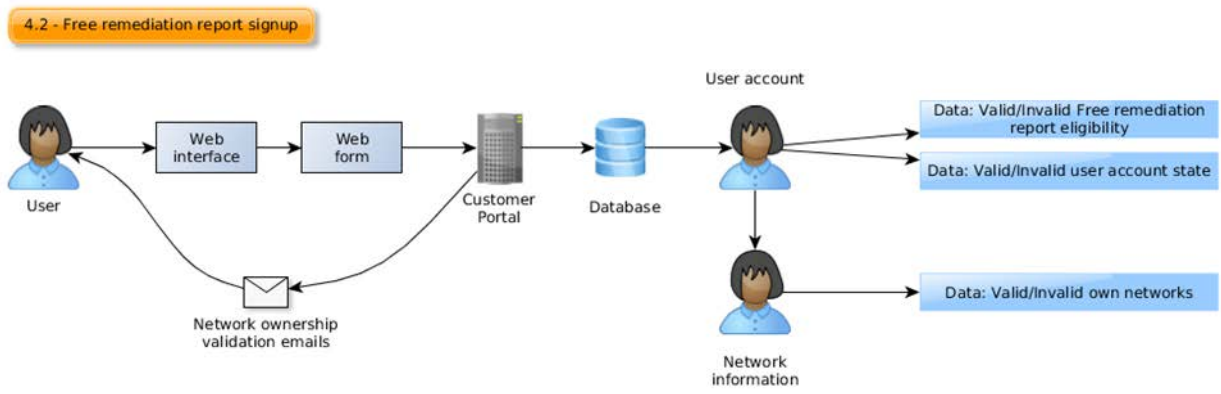
Actor state change: User -> Vetted User -> Report Recipient.

Data type(s): User account, User's network information (ASN/IP/CIDR), valid/invalid networks verified ownership, free remediation report eligibility.

Security controls: HTTPS/SSL website, Username/Password TLS email.

Data privacy: User data stored only on SISSDEN EU data centre web server, with privacy controls and opt-in/opt-out controls available. User's network information, email address and remediation report eligibility stored on SISSDEN EU data centre web server and Shadowserver Reporting System (covered under Privacy Shield). Opt-in/opt-out option will allow the Users to indicate if they want to have data about their networks collected or not. This implies that any data that allows identifying them will not be stored or distributed to third parties.

Diagram:



4.3 Customer Feedback System

Actor(s): User.

Interface type(s): Web interface, Email.

Interface location: Customer Portal.

Format(s): Web form, Support ticket, Email notification/response.

Interaction: A User with a validated User account visits customer SISSDEN website component, logs in successfully and provides feedback on their SISSDEN usage or experience in the form of a web based support/feedback ticket. Threaded tickets can be interacted with via the web interface or by email.

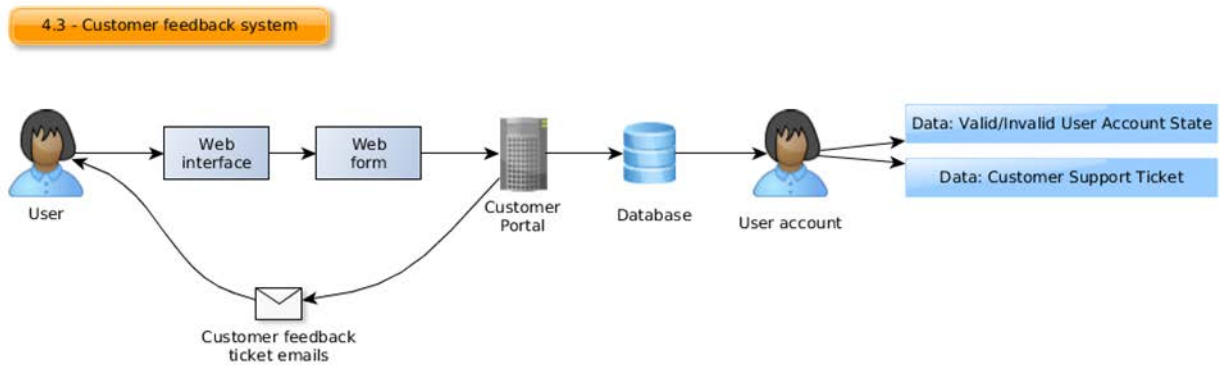
Actor state change: None.

Data type(s): Valid User account, Customer Feedback Ticket.

Security controls: HTTPS/SSL website, Username/password, TLS email.

Data privacy: User data stored only on SISSDEN EU data centre web server, with privacy controls and opt-in/opt-out controls available.

Diagram:



4.4 Shadowserver Reporting System

Actor(s): Report Recipient.

Interface type(s): E-mail, Web interface.

Interface location: SHAD external.

Format(s): Plain text email, download URL, CSV/XML data.

Interaction: Vetted User eligible for receiving remediation reports (Report Recipient) receives free of charge daily remediation reporting emails from SHAD's external Shadowserver Reporting System component. One or more emails contain either event data or a unique secure URL to download event data, with one email per data type (sissden-scans, sissden-spam, sissden-malware, etc). Data formats of the emails are defined on the public website, with examples provided in section 5. Event data format can be in CSV or XML (option to be chosen by Report Recipient).

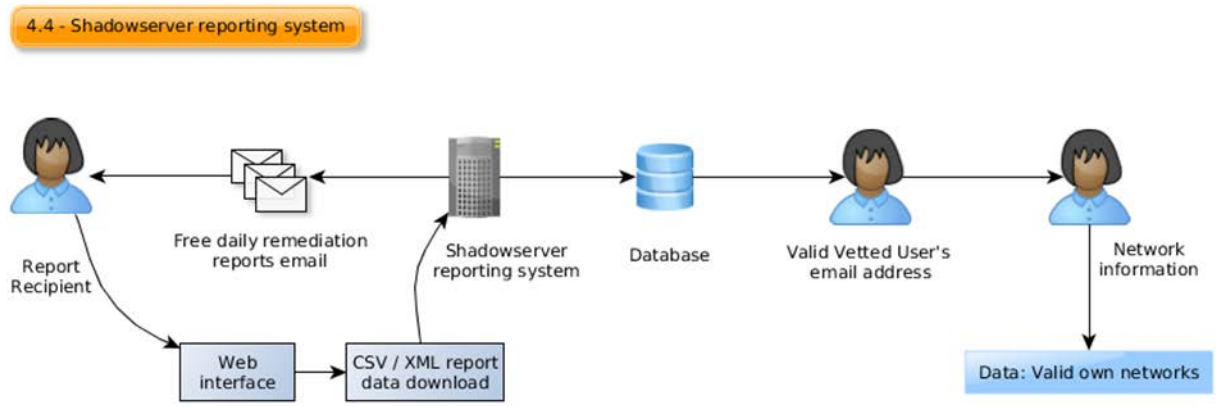
Actor state change: None.

Data type(s): Report Recipient's email address, Free remediation report eligibility, Free of charge daily remediation emails.

Security controls: TLS email, HTTPS/SSL website.

Data privacy: Report Recipient's network information, email address and remediation report eligibility stored on Shadowserver Reporting System (covered under Privacy Shield).

Diagram:



4.5 View Public Information about SISSDEN

Actor(s): Anonymous User.

Interface type(s): Web interface.

Interface location: Public website.

Format(s): Web form.

Interaction: Anonymous User views and navigates the public SISSDEN website pages. He/she can choose if cookies and tracking are enabled.

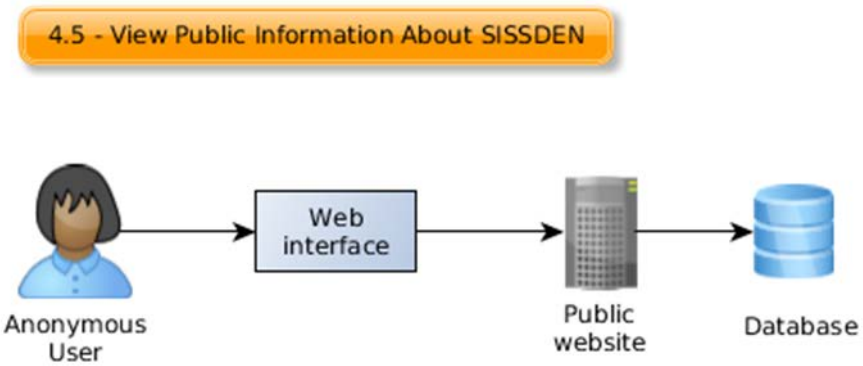
Actor state change: None.

Data type(s): Statistics will be collected to determine the number and type of users accessing the website.

Security controls: HTTPS/SSL website.

Data privacy: By default the Anonymous User will not be tracked (no cookies and no storing of IP address).

Diagram:



4.6 Subscribe/unsubscribe to news about SISSDEN

Actor(s): User or Anonymous User.

Interface type(s): Web interface.

Interface location: Customer Portal.

Format(s): Web form, Verification email.

Interaction: A User with a valid SISSDEN account logs in and visits a public SISSDEN website to access a Customer Portal component that allows signing up to SISSDEN news service consisting of periodic or on the fly email notifications containing important information on the project. The User's email address will be used to automatically send an email so that the User can confirm the subscription. The service can be parameterized so that the User can choose to receive news on the fly or combined in one mail each week or month.

Another option could be subscription by an Anonymous User, e.g. adding an email address to a subscription list or following SISSDEN's twitter feed.

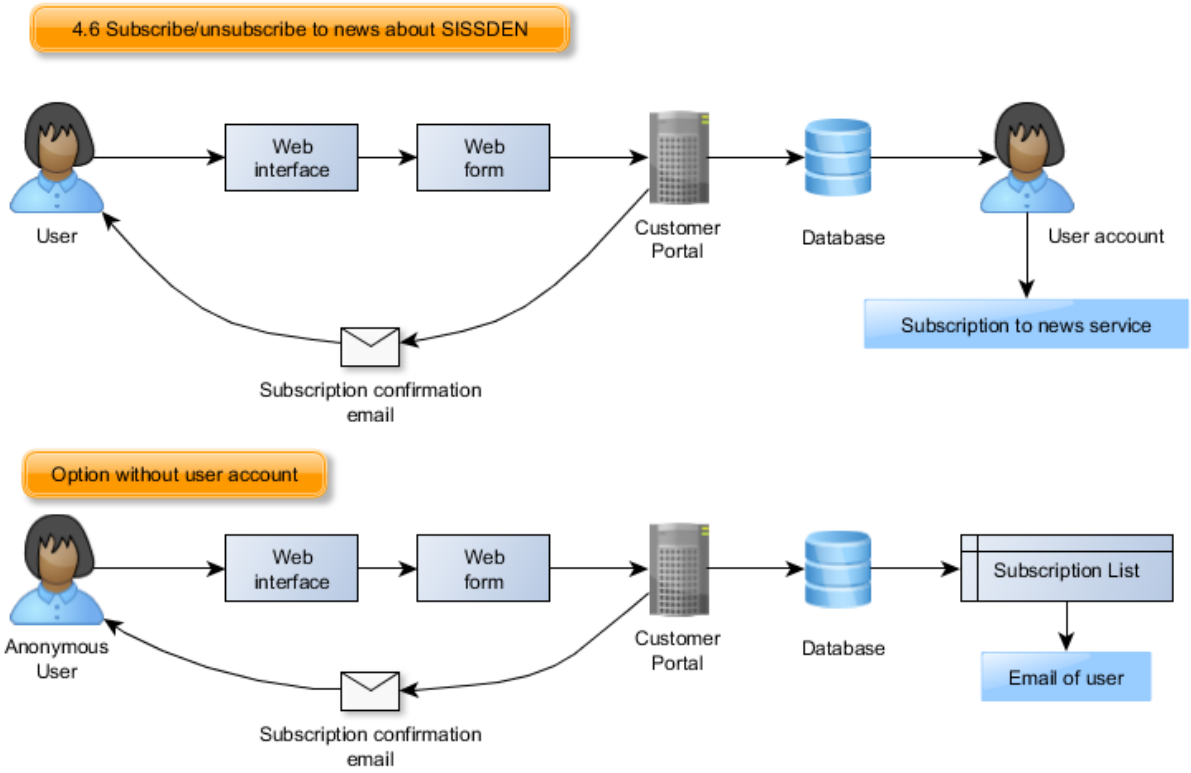
Actor state change: None.

Data type(s): User subscription to news service included in his/her profile or, in the case he/she doesn't have an account, in the news subscription list, Validation of User intent using an email with confirmation link.

Security controls: HTTPS/SSL website, Username/Password, TLS email.

Data privacy: User data stored only on SISSDEN EU data centre web server, with privacy controls and opt-out control to allow User to terminate the subscription.

Diagram:



4.7 Manage SISSDEN User account information

Actor(s): User.

Interface type(s): Web interface.

Interface location: Customer Portal.

Format(s): Web form.

Interaction: Interface that allows Users to manage their own account information.

A web interface within the Customer Portal hosted in the SISSDEN EU data centre will allow storing and updating the following information:

- User profile that contains all of the User's details and preferences. Examples include name, organisation, address, email, main IP address, date of account creation, salted encrypted password, date of last change, account preferences (tracking/cookies allowed or not, part or not of public membership list, keep or forget feedbacks, etc).
- Services subscriptions and preferences. Services include: news, remediation reports, etc. Preferences include: subscription or not to a service, date of subscription to a service, periodicity options, etc.
- Feedbacks: date of feedbacks, feedback tickets, actions taken (if any).

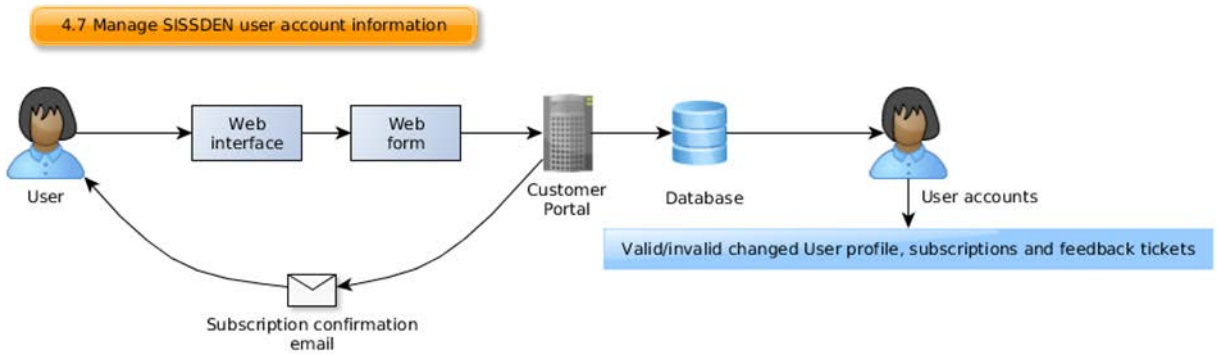
Actor state change: None.

Data type(s): User account, User profile data, User subscription data, User feedback data.

Security controls: HTTPS/SSL website, Username/Password, encrypted salted passwords.

Data privacy: User data stored only on SISSDEN EU data centre web server, with privacy controls and opt-in/opt-out controls available.

Diagram:



4.8 View high-level aggregated metrics

Actor(s): Anonymous User.

Interface type(s): Web interface.

Interface location: Metrics dashboard / Public website.

Format(s): Web page, Tables, Charts, Web form.

Interaction: An Anonymous User visits the public metrics dashboard, which displays tables and charts of key high-level aggregated metrics. By interacting with the web form, the Anonymous User can expand the view of these high-level aggregated metrics. For example, the metrics dashboard might show a table of the 10 countries with the most spam activity detected in the last month. In this example, the Anonymous User could interact with the web form to show the full list of countries, or to find where e.g. Australia ranks on the list.

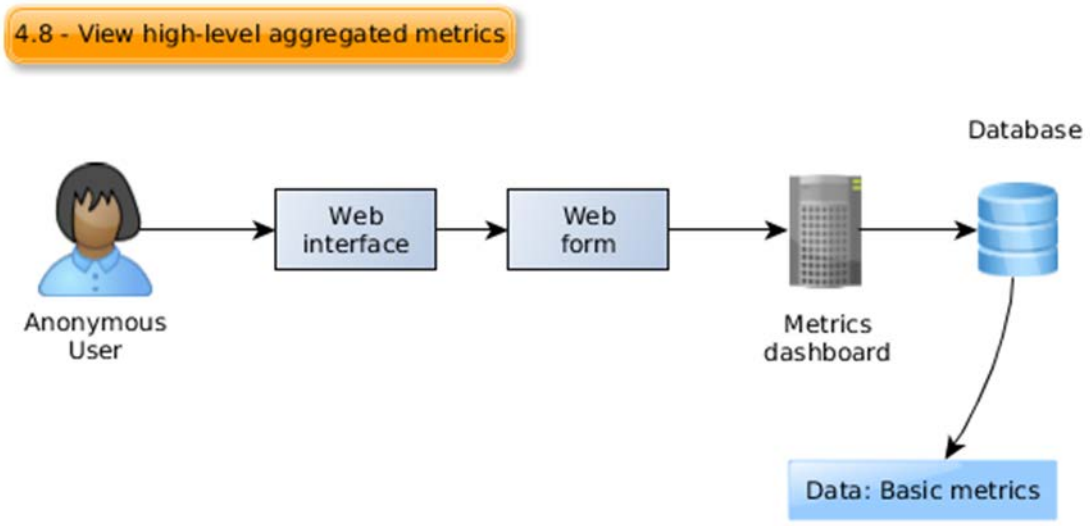
Actor state change: None.

Data type(s): Metrics (basic).

Security controls: HTTPS/SSL website.

Data privacy: By default, the Anonymous User will not be tracked by IP address and cookies will be used only to improve the UX of the site (PII will not be stored in cookies).

Diagram:



4.9 View more-detailed high-level aggregated metrics

Actor(s): Metrics Recipient.

Interface type(s): Web interface.

Interface location: Metrics dashboard.

Format(s): Web page, Tables, Charts, Web form, Web 2.0.

Interaction: A Vetted User who owns a network and is able to access metrics information (Metrics Recipient), visits a private metrics dashboard and interacts with the Web form or Web 2.0 application in order to view more advanced metrics and reports for one or more networks they have already been verified to own. If this network has been validated as being owned by the Vetted User (i.e. eligible for remediation reports), the Metrics Recipient will be able to view the available advanced metrics.

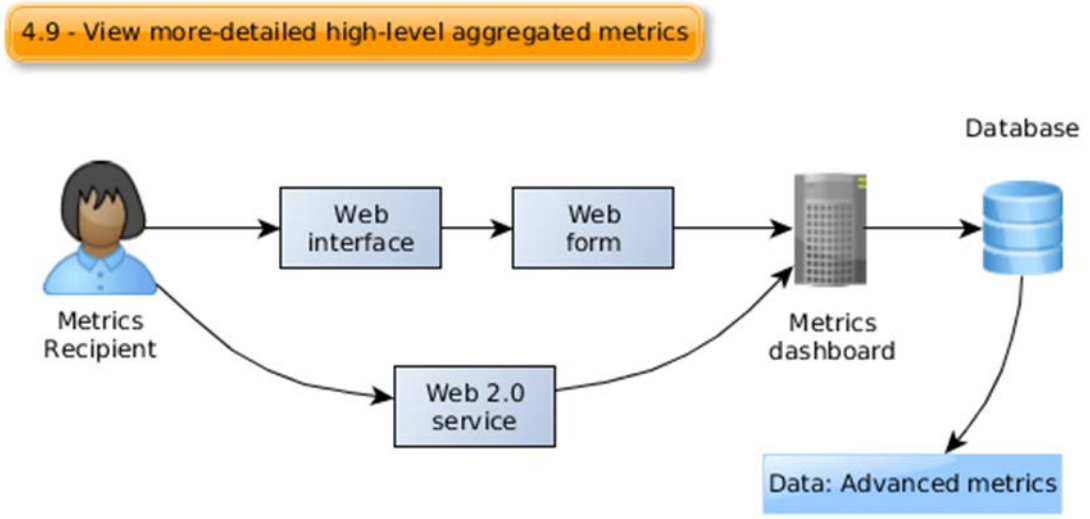
Actor state change: None.

Data type(s): Valid User account, Valid User's network information (ASN/IP/CIDR), Valid/invalid networks verified ownership, Metrics (advanced).

Security controls: HTTPS/SSL website, Username/Password.

Data privacy: User data stored only on SISSDEN EU data centre web server, with privacy controls and opt-in/opt-out controls available. Valid User's network information and remediation report eligibility stored on SISSDEN EU data centre web server and Shadowserver Reporting System (covered under Privacy Shield).

Diagram:



4.10 View and change data privacy settings

Actor(s): Vetted User.

Interface type(s): Website.

Interface location: Customer Portal.

Format(s): Web form.

Interaction: Vetted User with a valid SISSDEN account and verified network ownership modifies the accessibility of data pertaining to his networks. Users can also provide consent to share their personal data with third parties.

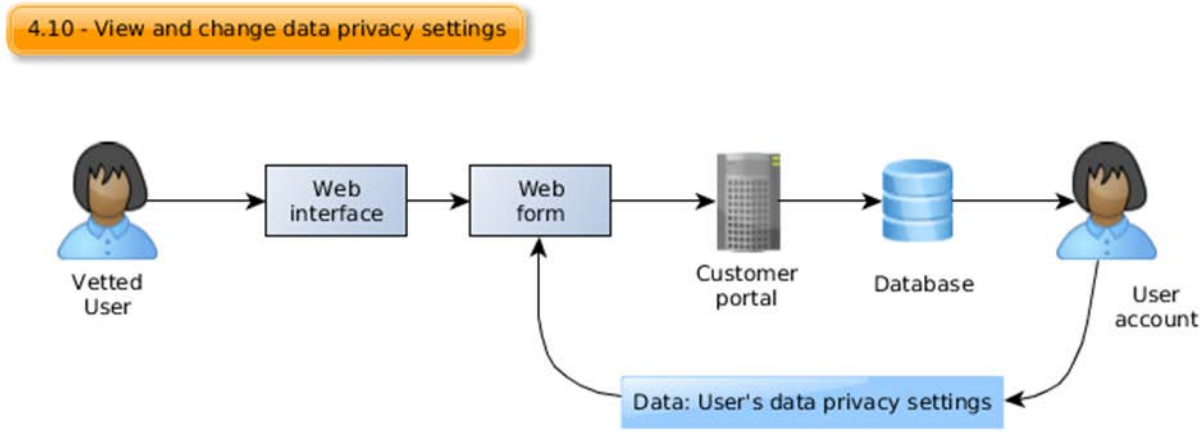
Actor state change: None.

Data type(s): Vetted User account, Vetted User's network information (ASN/IP/CIDR), Valid/invalid networks verified ownership, Free remediation report eligibility, Third Party Share Status.

Security controls: HTTPS/SSL website, Username/Password.

Data privacy: User data stored only on SISSDEN EU data centre web server, with privacy controls and opt-in/opt-out controls available.

Diagram:



4.11 View and change SISSDEN service opt-in/opt-out status

Actor(s): Vetted User.

Interface type(s): Web interface.

Interface location: Customer Website.

Format(s): Web form.

Interaction: A Vetted User visits the Customer Portal and wishes to exercise the right to opt out from SISSDEN storing or distributing any data that is collected about their network. They have already verified to SISSDEN that they own the network in question. When viewing the form, the opt-in/opt-out status is displayed for each network address space that the Vetted User has verified ownership of. The Vetted User chooses to opt out of a particular network address space. Once the User has opted out, they can similarly choose to opt back in.

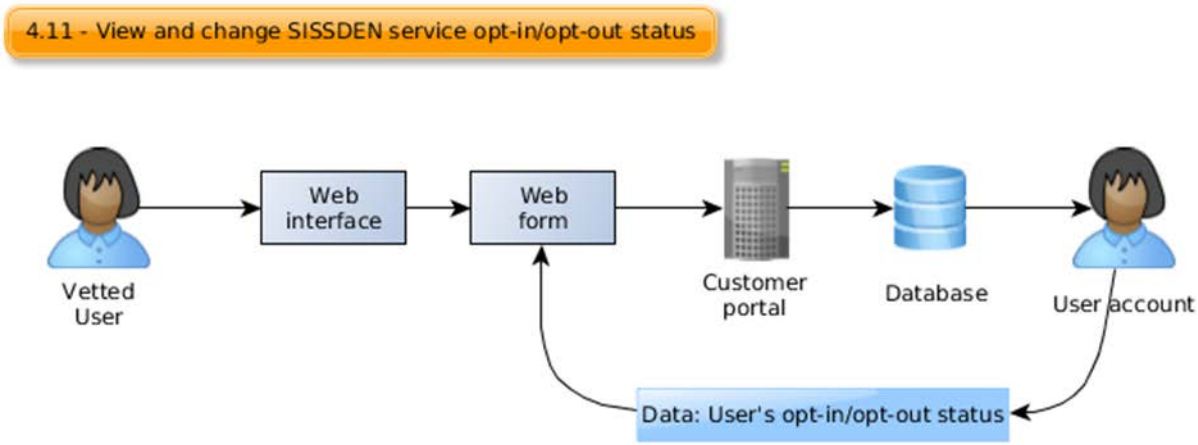
Actor state change: None.

Data type(s): Vetted User account, Vetted User's network information (ASN/IP/CIDR), Valid/invalid networks verified ownership.

Security controls: HTTPS/SSL website, Username/Password.

Data privacy: Vetted data stored only on SISSDEN EU data centre web server, with privacy controls and opt-in/opt-out controls available.

Diagram:



4.12 Sign up and request access to the SISSDEN curated reference data set

Actor(s): User.

Interface type(s): Web interface, Email.

Interface location: Customer Portal.

Format(s): Web form, Verification email.

Interaction: A User with a valid SISSDEN account visits the Customer Portal and accesses a page to request access to the curated reference data set. The User completes a form (which supplements data that is already known about the User account) and submits it. The information is submitted to the SISSDEN Partners and is responded to manually. Once eligibility to the curated reference data set is confirmed, an email is sent to the User to confirm that access has been granted.

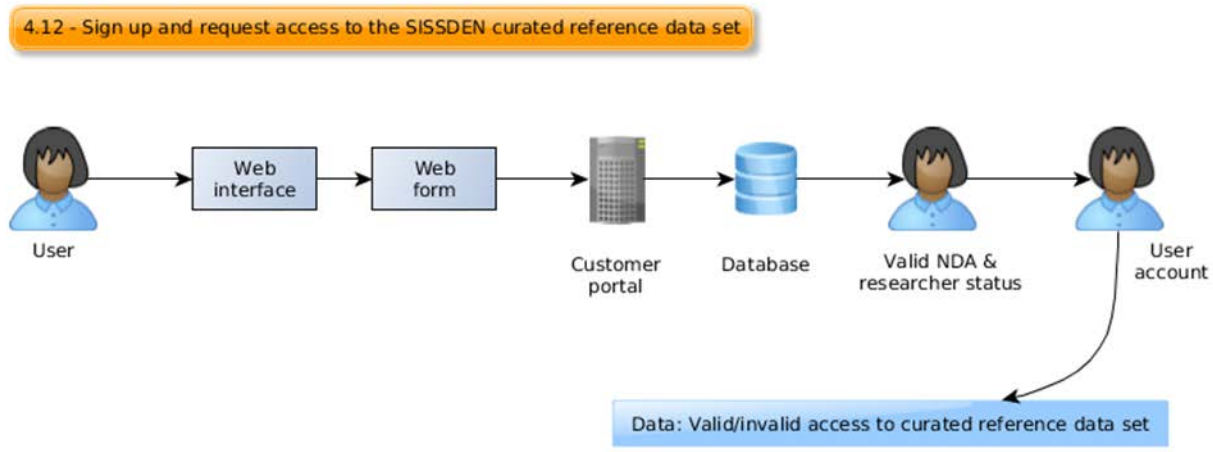
Actor state change: User -> Vetted User -> Approved Researcher.

Data type(s): Valid User account, Curated reference data set eligibility.

Security controls: HTTPS/SSL website, Username/Password, TLS email.

Data privacy: User data stored only on SISSDEN EU data centre web server, with privacy controls and opt-in/opt-out controls available.

Diagram:



4.13 Successfully vetted researchers access the SISSDEN curated reference data set

Actor(s): Approved Researcher.

Interface type(s): Web interface, Web service.

Interface location: Customer Portal.

Format(s): Web page, REST API.

Interaction: A Vetted User who has been authorised for access to the curated reference data set (Approved Researcher) visits the Customer Portal and accesses the page for the curated reference data set. This web page contains a description of the API for accessing files that make up the curated reference data set. The interface will allow data sets to be downloaded via, e.g., a RESTful API provided by the Customer Portal.

Actor state change: None.

Data type(s): Valid User account, Curated reference data set eligibility, Data feeds (e.g. in STIX 2.0, n6, custom JSON formats, raw formats).

Due to the scale of the data sets (expected to reach many terabytes) and the variation of data, it will not be possible to provide the entire data set in one unified format. Therefore, each data set will instead be delivered in the most suitable format for the data type. Where possible, data sets of events will be made available in the standard STIX 2.0 format over HTTPS. Additionally, data sets may be delivered in multiple alternative data formats (such as custom JSON or PCAP files), in order to ease integration for researchers unfamiliar with the standard.

Delivering data in standardized formats where possible will provide as much consistency between the reference data sets as possible (given the vastly different types of data), which increases the support for rapid, automated processing of data sets. This is in line with ENISA recommendations, which consistently encourage the use of standardized formats (such as STIX/TAXII) for threat intelligence and data exchange (e.g. recommended for CERTs[1] and malware analysts[2]).

[1] https://www.enisa.europa.eu/publications/detect-share-protect-solutions-for-improving-threat-data-exchange-among-certs

[2] https://www.enisa.europa.eu/topics/trainings-for-cybersecurity-specialists/online-training-material/documents/common-framework-for-artifact-analysis-activities-handbook

Examples of data sets that could be provided:

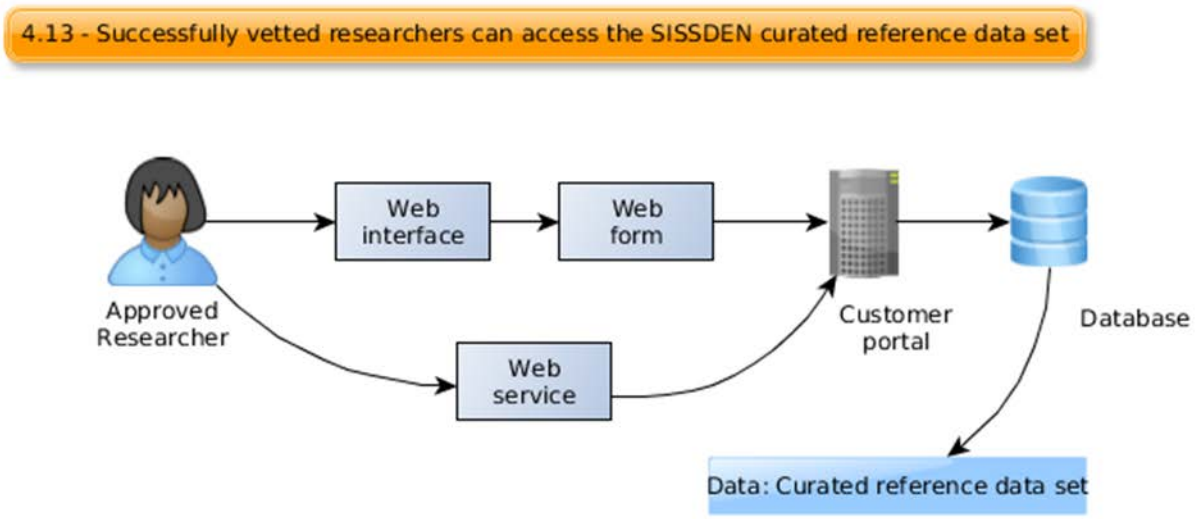
Data set	Contents of data set	Possible format
Sensor data	Raw data from all sensors	Netflow data in nfdump format
		Full packet captures in PCAP format
Honeypot data	Honeypot attack events	Custom JSON
		STIX 2.0
	Honeypot spam events	Custom JSON
		STIX 2.0
		Email messages in mbox format
	Honeypot malware events	Custom JSON
		STIX 2.0
	Honeypot malware binaries	Custom JSON + compressed archive
Darknet data	Raw data from darknet sensors	Full packet captures in PCAP format

In addition to data formats being chosen dependent on the nature of the data set, the same will be true of the level of concatenation of the data. For example, raw traffic data from the sensors will likely be split into one compressed file per hour of day of captured data, due to the large quantity of data involved.

Security controls: HTTPS/SSL website, Username/Password, HTTPS/SSL API with API key.

Data privacy: User data stored only on SISSDEN EU data centre web server, with privacy controls and opt-in/opt-out controls available. Data feeds stored only on SISSDEN EU data centre web server. Data feeds not anonymized, but restricted to vetted researchers only, as specified in D2.2.

Diagram:



5 SISSDEN free of charge victim remediation report/feed formats

The primary objective of SISSDEN is to deliver free of charge victim remediation reports/feeds to National CERTs, ISPs and netblock owners i.e. Report Recipients. These feeds will be generated based on the data collected by the SISSDEN sensor network of honeypots, along with data collected by other systems operated by SISSDEN consortium members and third-parties.

As explained in Section 4, data concerning malicious activity detected by SISSDEN will be included in Shadowserver's existing daily free of charge victim remediation feeds[3]. These will be made available to validated recipients (see Sections 4.2, 4.4 for an explanation of the sign up and reporting process). Data sets sent to recipients will be explicitly identified as SISSDEN by use of a hyphenated name sissden-* in the report source name.

SISSDEN will collect a wide variety of malicious activity types including:

- scanning activity detected by sensors (honeypots/darknets),
- brute force attempts to access sensors (honeypots),
- interactive login sessions to sensors (honeypots),
- spam e-mails (honeypots),
- malicious URLs identified as a result of observed activities (honeypots),
- malware binaries obtained as a result of observed activities (honeypots),
- additional URL/IP information extracted from obtained malware, including C&C information if any (via malware sandboxing),
- third-party sources that offer blacklists of IPs,
- … and many more over time.

We provide examples of potential data feeds delivered through SISSDEN in the next sections of this document. Note that these examples are provided in this document as initial samples, which will change over time once data starts being actively collected and new honeypots come online. Their inclusion here is intended to provide the reader with a better understanding of what SISSDEN Report Recipients will gain through subscribing to SISSDEN's remediation reports and illustrate the type of data collected.

[3] For a full overview of delivery mechanisms and report types currently supported by Shadowserver, see http://www.shadowserver.org/wiki/pmwiki.php/Services/Reports

5.1 Example reports based on the SISSDEN sensor network of honeypots

The SISSDEN network of honeypot sensors is the primary data collection building block of the SISSDEN platform. A number of initial sample formats of free daily remediation reports that will be generated based on the data collected by the SISSDEN sensor network is introduced below.

5.1.1 Observations of scanning activity

The following is an example report of scanning activity observed by the SISSDEN sensor network of honeypots (subject to change).

Field	**Description**
timestamp (first seen)	Timestamp the IP was first observed in UTC+0
timestamp (last seen)	Timestamp the IP was last observed in UTC+0
ip	IP of the scanner
asn	ASN of the scanner
geo	Geolocation of the scanner
seen by	Information how many SISSDEN sensors observed the same scanner
src port	Source port of the scan
dst port	Destination port of the scan
proto	Protocol
attack_name	Any attack details extracted by honeypots or supporting monitoring systems

5.1.2 Observations of brute force attack activity

The following is an example report of brute force attack activities (for example, against SSH services) observed by the sensor network of honeypots (subject to change).

Field	Description
timestamp (first seen)	Timestamp the IP was first observed in UTC+0
timestamp (last seen)	Timestamp the IP was last observed in UTC+0
ip	IP of the brute forcing device
asn	ASN of the brute forcing device
geo	Geolocation of the brute forcing device
seen by	Information how many SISSDEN sensors observed the same brute forcing device
service	Service being attacked, typically ssh, telnet...
src port	Source port used in the attack
dst port	Destination port of the service being attacked
count	How many brute force attempts were observed

5.1.3 Observations of malware activity

The following is an example report of malware that was observed by the SISSDEN sensor network of honeypots (subject to change).

Field	Description
timestamp (first seen)	Timestamp the URL was first observed in UTC+0
timestamp (last seen)	Timestamp the URL was last observed in UTC+0
md5hash	MD5 of the downloaded binary
url	URL where binary was downloaded from
url_ip	IP of the URL
url_asn	ASN of the URL
url_geo	Geolocation of the URL

5.1.4 Observations of spam activity

The following is an example report of spam messages observed by the SISSDEN sensor network of honeypots (subject to change).

Field	Description
timestamp	Timestamp of the message in UTC+0
url	URL that was extracted from a Spam message
host	Hostname of the URL location
url_ip	IP of the URL
asn	ASN where the IP resides
geo	Country location of the IP
subject	Subject of the Spam message
src_ip	IP address of the Spam relay that delivered the message (last hop)
src_ip_asn	ASN of the relay IP
src_ip_geo	Country location of the Spam relay
relay method	SMTP, SOCKS etc.

5.2 Example reports based on SISSDEN Partner systems

Multiple SISSDEN Partners operate a number of supporting data collection systems that SISSDEN will also regularly ingest data from. This includes sandboxing platforms, networks of honeypot sensors and in-house threat intelligence solutions.

5.2.1 Observations of C&C activity

The following is an example report of Command and Control servers (C&C) extracted from malware from SISSDEN Partner sandbox systems (subject to change).

Field	Description
timestamp (first seen)	Timestamp the C&C was first observed in UTC+0
timestamp (last seen)	Timestamp the C&C was last observed in UTC+0
url	URL of C&C (if any)
ip	IP of C&C
asn	ASN of the C&C
geo	Geolocation of the C&C
label	Any threat label that was applied (malware family name, threat actor, etc.)

5.3 Example reports based on third party sources

The various SISSDEN system components are complemented through data generated by third parties (primarily public sources, outside of the SISSDEN consortium).

5.3.1 Observations of blacklisted devices

The following is an example report that aggregates information from third party sources feeding SISSDEN (subject to change).

Field	Description
timestamp	Timestamp of tracked event in UTC+0
ip	IP address of device in question
hostname	Reverse lookup of IP in question
source	Blacklist source
reason	Given reason of blacklisting by the source
asn	ASN of the IP
geo	Geolocation of the IP

6 Data protection

As already established in deliverables D1.4 “Guidelines for data handling and data sharing with partners” and D2.2 “Preliminary legal requirements”, the data collected and processed within SISSDEN will likely include personally identifiable information (PII). Therefore, it is the consortium’s responsibility to ensure proper security of the data. This chapter discusses this aspect of the project. The first section presents the approach taken to regulate the access of different user groups to interfaces and data published by SISSDEN. The second section discusses the issues of data privacy and anonymization.

6.1 Data access control approach

Due to the amount of potentially sensitive data stored and processed in the system, proper access control is essential. Some parts of the system are public facing and do not require any access restrictions, while access to other elements requires careful vetting of potential users.

In general, the approach taken in the project follows the industry standard RBAC (Role-Based Access Control) model[4], as described in the standard ANSI/INCITS 359-2012[5]. This approach is fairly common in modern complex systems and enables scalable administration of large sets of users and resources without sacrificing granularity. The list of actors described earlier in this document clearly maps to an initial set of necessary roles and permissions.

The RBAC approach is sufficient to regulate access to different services of the SISSDEN system. However, on the level of individual services, more fine-grained access control will be needed to preserve the privacy requirements. Most importantly, following the findings of the deliverables D1.4 “Guidelines for data handling and data sharing with partners” and D2.2 “Preliminary legal requirements”, sharing of data that may contain personally identifiable information with external partners requires sufficient justification and must demonstrate significant public benefits. In this case a more complex access model is needed, combining the role-based approach with attribute-based access control.

The general category of Vetted Users covers these cases. In general, the vetting process verifies the justification for access to a particular data set and results in the granting of an appropriate role and assignment of the proper attributes (e.g. AS number, network CIDR, etc). The actual process depends on the type of data requested. A more thorough identity check and additional communication may be required to verify the authenticity of a request and the validity of the requested scope. Also, depending on the requested data, the vetting process must include the verification of the data receiver’s ability to conform to data processing requirements as enforced by European law (see next section for discussion of the privacy concerns within SISSDEN). In case of major privacy concerns, the process may need to be much more formal and require signing an agreement clarifying the data processing restrictions. This applies mostly when the requested data is not actionable information pertaining to the network the User is responsible for. The details of the vetting process will be elaborated in future deliverables.

[4] Ferraiolo, D.F. and Kuhn, D.R., 1992. Role Based Access Control. In: Proceeding of the 15th National Conf. On Computer Security Conference, Elsevier Advanced Technology Publications, Oxford, UK, October 13-16, pp. 554-563.

[5] Formal compliance with this standard is not an explicit goal of this project, but the general approach adopted is the same.

Whenever sensitive data is to be shared with such a user, apart from the verification of the role itself, the role attributes must be checked as to whether they correspond to the characteristics of the data in question. This fine-grained approach ensures that e.g. a network owner will only see non-anonymized logged traffic from honeypots if it originates from or targets his own network.

The detailed implementation of access controls is a part of each public interface's design and will be documented in the private internal SISSDEN technical architecture document (updated versions of D3.3, to be delivered as part of future deliverables).

For authentication of a user, a standard username/password approach is deemed sufficient for the security needs of the project. The communication should also be protected using available industry-standard approaches, such as e.g. SSL/TLS. The choice of applicable technologies may depend on the requirements of a specific interface. The initial decisions are documented in the descriptions of individual interfaces, but may be subject to changes – both as a result of the progress of the project itself and of the evolution of state of the art.

6.2 Privacy and Anonymization

SISSDEN collects personal identifiable information such as IP and e-mail addresses during the collection and processing of data obtained from the interactions with honeypots, the analysis of samples of malware in sandboxes or as part of the data flows recorded in darknets and other deployed data analysis probes.

While current data protection legislation does not allow the indiscriminate disclosure of personally identifiable information to the public at large, the legislation permits the processing of such information so as to strike the right balance between personal integrity protection and the protection of Internet services, the fight against cybercrime or the remediation of compromised servers and services.

SISSDEN operates its own sensor network, and the raw data collected is strictly the result of the network interactions with SISSDEN sensors (i.e. unsolicited probes and attacks against them). More data is obtained as a result of the processing of the malware samples run in controlled sandbox environments. Although the processing of such malware samples does not take place in the same location that they are originally collected, SISSDEN has designed the system to have the biggest guarantees that all the data collected was intentionally aimed at our sensor network with solely malicious intent.

SISSDEN will disclose such personally identifiable information to entities affected by the attacks (victims) and to those legally appointed to work in the remediation of the related malicious activity (National CERTs, Law Enforcement Agencies and network owners). When the nature of the Internet threat has the potential to affect the public at large, SISSDEN can make such information available to wider communities.

SISSDEN will also deliver added value services such as the "metrics service", where no personally identifiable information will be disclosed. To achieve anonymization of personal data, the service will aggregate sensitive and personal data and provide indicators of malicious activity within entire networks, autonomous systems or countries.

7 Annex 1 - n6: REST API v0.17

This Annex describes the basic search interface for NASK's background n6 platform, which will be used in and extended during the project. Since the SISSDEN system and the n6 platform are still under development, parts of this specification may be subject to change.

7.1 Overview

n6 uses an event-based data model for representation of all types of security information. Each event is natively represented as a JSON object with a set of mandatory and optional attributes (see "Event attributes" section below).

The REST API is available over TLS with mandatory authentication via client certificates. ABNF syntax of the generic URI scheme:

```
"https://" server "/" resource "." format "?" query
```

where

- server is a fully-qualified domain name of the API server,
- resource is used to identify the desired scope of the data received, for the global dataset it must be set to search/events,
- format is the requested format, can be json, sjson, csv or iodef,
- query defines which events should be served, described in the next section.

7.2 Query

A query consists of a list of conditions on values of selected attributes. Query syntax in ABNF:

```
query = arg *( "&" arg )

arg = name [ "=" [ value ] ]

name = plain

value = plain / set

set =  plain *( "," plain )
```

where plain is a percent-encoded (RFC 3986) character string. "Safe" characters that do not require encoding: ALPHA / DIGIT / "-" / "." / "_" / "~", others must be encoded.

name corresponds to the name of the event attribute. Any attribute of string or numeric type can be used in queries. The name can be followed by the equals character and the requested value of the attribute. If no value is given an empty string is assumed. Multiple values can be specified at the same time by separating them with commas or, alternatively, by repeating the attribute with different values.

Examples of complete URIs containing queries:

https://FQDN/RESOURCE.json?ip=10.0.0.1&modified.min=2016-01-01T00:00:00Z

https://FQDN/RESOURCE.json?ip=10.0.0.1&modified.min=2016-01-01T00:00:00Z&opt.primary

https://FQDN/RESOURCE.json?name=%27%25xxx%27%3D

https://FQDN/RESOURCE.json?name=malware1,malware2

https://FQDN/RESOURCE.json?name=malware1&name=malware2

7.3 Response

n6 uses standard HTTP status codes: 200 (success), 400 (incorrect query), 404 (incorrect resource), 403 (no permission), 405 (incorrect HTTP method), 500 (server error).

Contents of the reply depend on the format requested; for JSON it is a single array where elements correspond to individual events. Each event is represented as a single JSON object with elements (keys) defined in the next section.

For large responses we recommend using “streamed” JSON variant (SJSON) which consists of concatenated top-level objects delimited by newlines (line feed, 0x10 ASCII). Each top-level object is represented in a single line (no pretty-print), which allows incremental parsing of results. Otherwise this format is identical with plain JSON.

In case of error just a text description is returned.

7.4 Event attributes

All attributes supported by the current version of n6 are listed below. [mandatory] denotes keys that must be present in their parent objects, by default all elements are optional. Element types are noted in brackets.

- action [string]: Action taken by malware, e.g. redirect, screen grab.
- address [array of objects]: Object containing IP address related to the threat. For malicious websites - A records in DNS, for connections to sinkhole and scanning hosts - source IP address. Elements of child objects:
    - ip [string] [mandatory]: IPv4 address in dot-decimal notation.
    - ipv6 [string] [mandatory]: IPv6 address in the hexadecimal notation. `ipv6` and `ip` are mutually exclusive - no more than a single address can be an element of the same object.
    - cc [string]: Country code (ISO 3166-1 alpha-2).
    - asn [integer]: Autonomous system number (without “AS” prefix).
    - dir [string]: Role of the address in terms of the direction of the network flow in layers 3 or 4. Possible values: `src` (address is the source of the flow) / `dst` (address is the destination of the flow).
    - rdns [string]: PTR record of the `.in-addr-arpa` domain associated with the IP address (without the terminal dot).

- adip [string]: Anonymized destination address (see dip) in dot-decimal address without prefix, e.g. x.184.216.119.
- category [string] [mandatory]: Category of the event. Possible values:
  - amplifier: hosts that can be used in amplification attacks (DoS),
  - bots: infected machines,
  - backdoor: addresses of web shells or other types of backdoors installed on compromised servers,
  - cnc: botnet controllers,
  - deface: compromised websites,
  - dns-query: DNS queries and answers (no determination on legitimacy / maliciousness),
  - dos-attacker: (distributed) denial-of-service attacks - details related to sources,
  - dos-victim: (distributed) denial-of-service attacks - details related to victims,
  - flow: network traffic in layer 3 (no determination on legitimacy / maliciousness),
  - flow-anomaly: anomalous network activity (not necessarily malicious),
  - fraud: activities and entities related to financial fraud,
  - leak: leaked credentials or personal data,
  - malurl: malicious URLs (details about web servers infecting Users),
  - malware-action: actions that malware is configured to make on infected machines,
  - phish: phishing campaigns (similar to malurl),
  - proxy: open proxy servers,
  - sandbox-url: URLs contacted by malware,
  - scam: sites offering fake content,
  - scanning: hosts performing port scanning,
  - server-exploit: attackers actively attempting to exploit servers,
  - spam: hosts sending spam,
  - spam-url: addresses found in spam,
  - tor: Tor network nodes,
  - webinject: injects used by banking trojans,
  - vulnerable: addresses of vulnerable devices or services,
  - other: other activities not included above.
- confidence [string] [mandatory]: Level of trust that the information is accurate. Possible values: low / medium / high.

- count [integer]: Connection (or other activity) count related to the event (applicable only to events resulting from aggregated data).
- dip [string]: Destination IP address (e.g. sinkhole, honeypot) in dot-decimal notation. Does not apply to addresses of malicious websites.
- dport [integer]: Destination port used in TCP or UDP communication.
- email [string]: Email address associated with the threat (e.g. source of spam, victim of a data leak).
- expires [string]: Time until the blacklist entry is considered valid.
- fqdn [string]: Fully-qualified domain name related to the threat. For malicious websites - domain in the URL; for bots and scanners - destination domain.
- iban [string]: International Bank Account Number associated with fraudulent activity.
- id [string] [mandatory]: System-wide unique event identifier.
- injects [array of objects]: Objects describing a set of injects performed by banking trojans when a User loads a targeted website (see `url_pattern`). Exact structure of injects is dependent on malware family and not specified at this time.
- md5 [string]: MD5 hash of the binary file related to the event.
- name [string]: Category-dependent name of the threat, e.g. `virut`, `SSH Scan`.
- origin [string]: Method used to obtain the data. Possible values:
  - `c2`: direct botnet controller observation,
  - `dropzone`: botnet dropzone observation,
  - `proxy`: monitoring traffic on a proxy server,
  - `p2p-crawler`: active crawl of a peer-to-peer botnet,
  - `p2p-drone`: passive listening to traffic in a peer-to-peer botnet,
  - `sinkhole`: data obtained from sinkhole,
  - `sandbox`: results from behavioural analysis,
  - `honeypot`: interaction with honeypots, both client and server-side,
  - `darknet`: monitoring of traffic collected by darknet,
  - `av`: reports from anti-virus systems,
  - `ids`: reports from intrusion detection and prevention systems,
  - `waf`: reports from web application firewalls.
- proto [string]: Protocol used on top of the network layer: `tcp` / `udp` / `icmp`.
- restriction [string] [mandatory]: Classification level, possible values: `internal` / `need-to-know` / `public`.
- sha1 [string]: SHA1 hash of the binary file related to the event.
- source [string] [mandatory]: Source (producer) of the event.

- sport [integer]: Source port used in TCP or UDP communication.
- phone [string]: Telephone number, national or international. Consists of numbers, optionally prefixed by the plus symbol.
- registrar [string]: Name of the domain registrar.
- status [string]: Blacklist entry status. Possible values:
  - active: item currently in the list,
  - delisted: item marked as inactive by an external source,
  - expired: item is considered no longer active but might be still present in an external blacklist,
  - replaced: some characteristics of an entry have changed and are represented as a new event (e.g. IP address change).
- replaces [string]: Identifier (id) of the event that was superseded by the current one. Specific to blacklists.
- target [string]: Organization or brand that is target of the attack (applicable to phishing).
- time [string] [mandatory]: Time of the occurrence (not time of reporting), format defined in RFC 3339.
- until [string]: Time of the last activity related to the event (applicable only to events resulting from aggregated data).
- url [string]: URL related to the event, format defined in RFC 3986.
- url_pattern [string]: Wildcard pattern or regular expression triggering injects used by banking trojans.
- Username [string]: Local identifier (login) of the affected User.
- x509fp_sha1 [string]: SHA-1 fingerprint of an SSL certificate in hexadecimal format.

Attributes not listed above might appear in the results to represent source-specific data elements. Syntax and semantics of such attributes are not defined in this document.

Additionally, the following pseudo-attributes can be used in queries for specifying wider search criteria:

- url.sub [string]: Substring in the url attribute.
- fqdn.sub [string]: Substring in the fqdn attribute.
- ip.net [string]: IPv4 network in CIDR notation, e.g. 10.0.0.0/8.
- ipv6.net [string]: IPv6 network, e.g. 2001:DB8::/32.

A special class of pseudo-attributes are ones that refer to time ranges. Names of these attributes consists of two parts, where the first on defines data that is being queried:

- active [string]: Refers to time and expires attributes: both are used for comparison and if either of them falls into the requested range, the whole criterion matches. E.g.

active.min=2014-10-04 would select events that either started after 2014-10-04 or started earlier but were still active after that date.

- modified [string]: Time when data was made available through the API (e.g. time when the record was inserted into the internal database) or when content of an existing event has changed.
- time [string]: Refers to the real time attribute.

The second part of the name of a pseudo-attribute consists of a one of the following operators:

- .min value is no earlier than the right-hand argument (inclusive),
- .max value is no later than the right-hand argument (inclusive),
- .until value is smaller than right-hand argument (exclusive).

There is also a special pseudo-attribute available that makes it possible to limit queried information to the original "primary data" from the event source, i.e., to exclude any information that has been added or inferred by the n6 system, especially any data from DNS/GeoIP queries made by n6:

- opt.primary [flag].

The above [flag] indicator means that possible values of opt.primary are:

- to turn on the feature (to obtain "primary data" only): yes, y, true, t, on, 1 or empty string (just &opt.primary can be appended to a query);
- to turn off the feature (the default behaviour, as without opt.primary): no, n, false, f, off, 0.

7.5 Sample document in n6 format

Plain JSON format:


```
[
    {
        "address": [
            {
                "ip": "195.187.240.100",
                "cc": "PL",
                "asn": 12824
            }
        ],
        "adip": "x.2.137.140",
        "category": "bots",
        "confidence": "medium",
        "count": 18,
        "dport": 80,
        "fqdn": "example.com",
        "id": "26c8fd5097251dd15dc8431b267c65cf",
        "name": "B58-DGA2",
        "origin": "sinkhole",
        "proto": tcp,
        "source": "b",
        "sport": 51869,
        "time": "2013-09-18T15:35:32",
        "until": "2013-09-18T19:00:00"
    },
    {
        "address": [
            {
                "cc": "PL",
                "ip": "108.162.201.25",
                "asn": 8308
            }
        ],
        "category": "malurl",
        "confidence": "low",
        "fqdn": "www.unknown-malware.eu",
        "id": "e1a53668ec9a2fe85974086815559868",
        "origin": "honeypot",
        "source": "m",
        "time": "2013-09-18T11:06:10",
        "url": "http://www.unknown-malware.eu/index.html?q=1"
    }
]
```


The same document in SJSON (lines truncated for readability):


```
{"address":[{"ip":"195.187.240.100","cc":"PL","asn":12824}],"adip" ...
{"address":[{"cc":"PL","ip":"108.162.201.25","asn":8308}],"categor …
```

8 Annex 2 - SISSDEN Architecture Diagrams

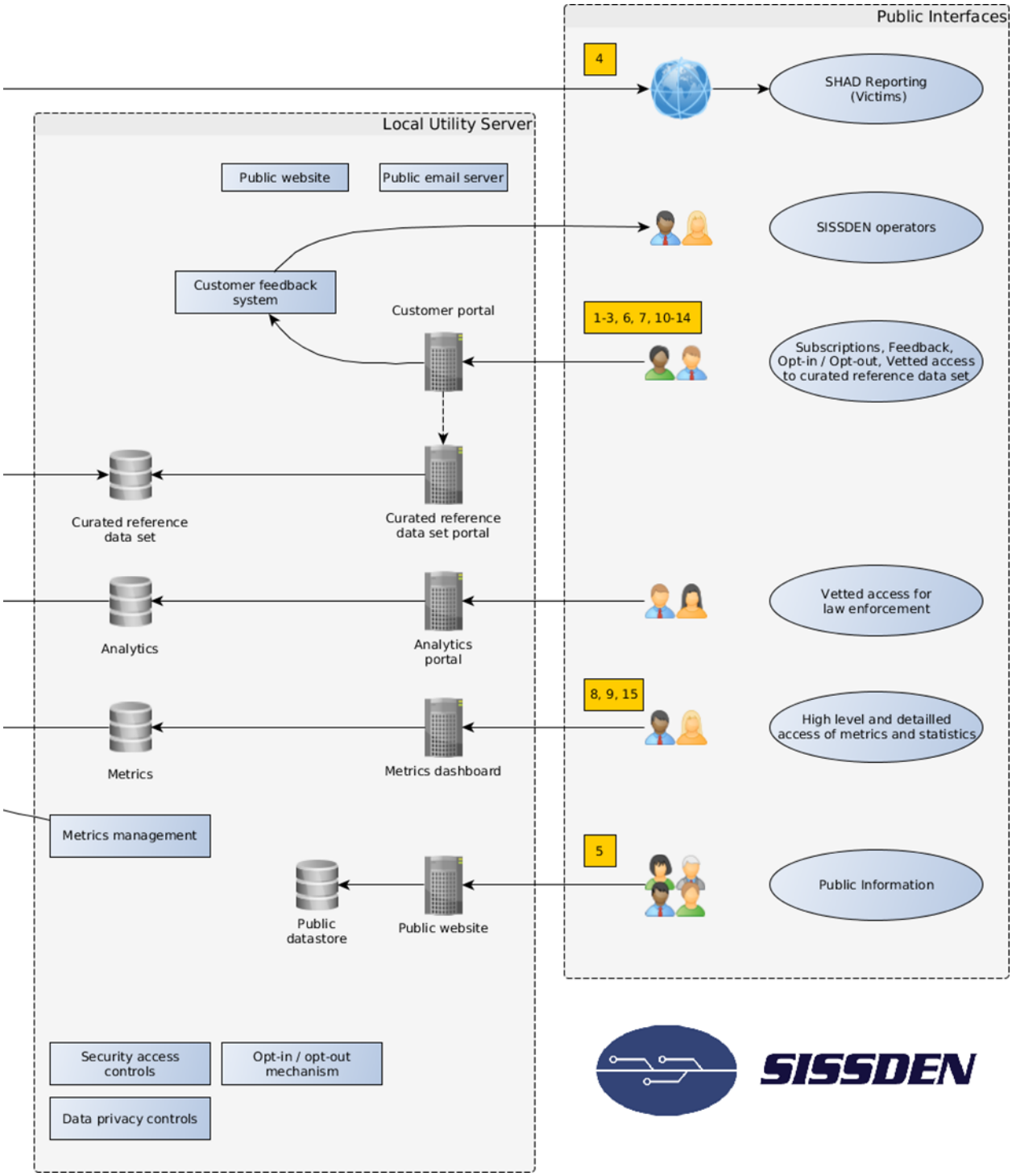


Figure 8.1: Schematic Frontend Architecture

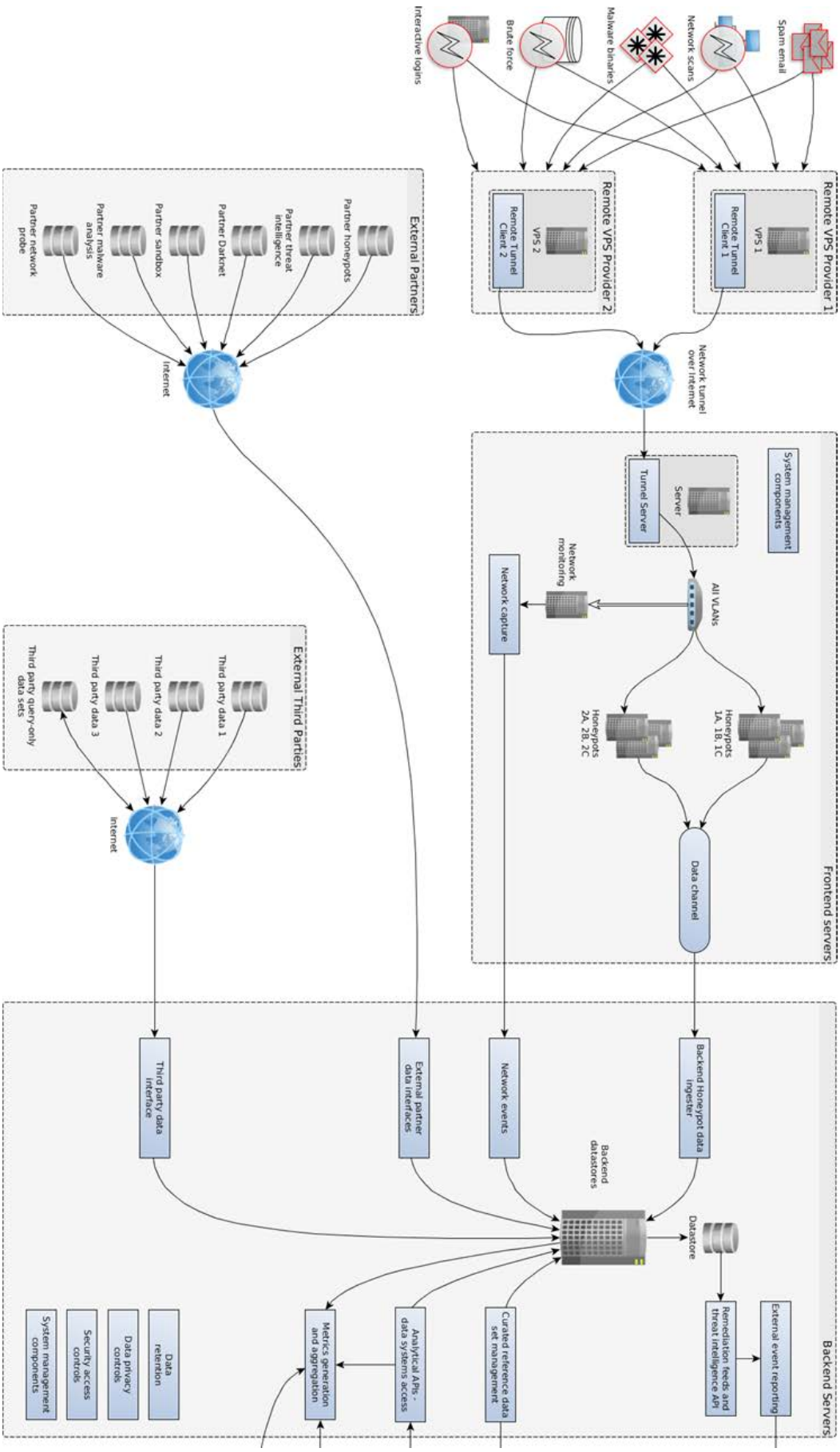


Figure 8.2: Schematic Backend Architecture